HDR-VQM: An Objective Quality Measure for High Dynamic Range Video

Manish Narwaria, Matthieu Perreira da Silva, Patrick Le Callet

HDR-VQM: An Objective Quality Measure for High Dynamic Range Video

Manish Narwaria, Matthieu Perreira Da Silva, Patrick Le Callet

Abstract

High Dynamic Range (HDR) signals fundamentally differ from the traditional low dynamic range (LDR) ones in that pixels are related (proportional) to the physical luminance in the scene (i.e. scene-referred). For that reason, the existing LDR video quality measurement methods may not be directly used for assessing quality in HDR videos. To address that, we present an objective HDR video quality measure (HDR-VQM) based on signal pre-processing, transformation, and subsequent frequency based decomposition. Video quality is then computed based on a spatio-temporal analysis that relates to human eye fixation behavior during video viewing. Consequently, the proposed method does not involve expensive computations related to explicit motion analysis in the HDR video signal, and is therefore computationally tractable. We also verified its prediction performance on a comprehensive, in-house subjective HDR video database with 90 sequences, and it was found to be better than some of the existing methods in terms of correlation with subjective scores (for both across sequence and per sequence cases). A software implementation of the proposed scheme is also made publicly available for free download and use.

Index Terms

High Dynamic Range (HDR) video quality, objective quality, spatio-temporal analysis.

I. Introduction

The advent of better technologies in the field of visual signal capture and processing has fueled a paradigm shift in todays' multimedia communication systems. As a result, the notion

The authors are with IRCCyN/IVC group, University of Nantes, 44306, France e-mail: (manish.narwaria@univ-nantes.fr, matthieu.perreiradasilva@univ-nantes.fr, patrick.lecallet@univ-nantes.fr).

of network-centric Quality of Service (QoS) in multimedia systems is being extended by relying on the concept of Quality of Experience (QoE) [1]. In this quest of increasing the immersive video experience and the overall QoE of the end user, newer technologies such as 3D, ultra high definition (UHD) and, more recently, High Dynamic Range (HDR) imaging have gained prominence within the multimedia signal processing community. HDR in particular has attracted attention since it in a way revisits the way we capture and display natural scenes. This is motivated by the fact that natural scenes often exhibit large ranges of illumination values. However, such high luminance values often exceed the capabilities of the traditional low dynamic range (LDR) capturing and display devices. Consequently, it is not possible to properly expose the dark and the bright areas simultaneously in one image (or video) during capture. This may lead to over-exposure (saturated pixels that are fully white) and/or under-exposure (very dark or noisy pixels as sensor's response falls below its noise threshold). In both cases, visual information is either lost or altered. HDR imaging focuses on minimizing such losses and therefore aims at improving the quality of the displayed pixels by incorporating higher contrast and luminance.

As a result, HDR imaging has attracted attention from both academia and industry, and there has been interest and effort to develop tools/algorithms for HDR video processing [2]. For instance, there have been recent efforts within the Moving Picture Experts Group (MPEG) for extending High Efficiency Video Coding (HEVC) to HDR. Likewise, the JPEG has announced extensions that will feature the original JPEG standard with support for HDR image compression. However, there is lack of such effort to quantify and measure the impact of such tools on HDR video quality using both subjective and objective approaches. The issue assumes further significance given that most of the existing objective methods may not be directly applicable for HDR quality estimation [3], [4] and [5] (note that these studies only deal with HDR images and not video). It is therefore important to develop objective methods for HDR video quality measurement and benchmark their performance against subjective ground truth.

With regards to visual quality measurement, both subjective and objective approaches can be used. The former involves the use of human subjects to judge and rate the quality of the test stimuli. With appropriate laboratory conditions and a sufficiently large subject panel, it remains the most accurate method. The latter quality assessment method employs a computational (mathematical) model to provide estimates of the subjective video quality. While such objective models may not mimic subjective opinions accurately in a general scenario, they can

be reasonably effective in specific conditions/applications. Hence, they can be an important tool towards automating the testing and standardization of HDR video processing algorithms such as HDR video compression, post-processing, inverse video tone mapping etc. especially when subjective tests may not be feasible. In light of this, we present a computationally tractable HDR video quality estimation method based on HDR signal transformation and subsequent analysis of spatio-temporal segments, and also verify its prediction performance based on a test bed of 90 subjectively rated compressed HDR video sequences. To the best of our knowledge, our study is amongst the first few efforts towards the design and verification of an objective quality measurement method for HDR video, and is therefore of interest to the video signal processing community both from subjective and objective quality view points.

II. BACKGROUND

Humans perceive the outside visual world through the interaction between luminance (measured in candela per square meter cd/m^2) and the eyes. Luminance first passes through the cornea, a transparent membrane. Then it enters the pupil, an aperture that is modified by the iris, a muscular diaphragm. Subsequently, light is refracted by the lens and hits the photoreceptors in the retina. There are two types of photoreceptors: cones and rods. The cones are located mostly in the fovea. They are more sensitive at luminance levels between 10^{-2} cd/m^2 to 10^8 cd/m^2 (referred to as the photopic or daylight vision) [6]. Further, color vision is due to three types of cones: short, middle and long wavelength cones. The rods, on the other hand, are sensitive at luminance levels between 10^{-6} cd/m^2 to 10 cd/m^2 (scotopic or night vision). The rods are more sensitive than cones but do not provide color vision.

Pertaining to the luminance levels found in the real world, direct sunlight at noon can be of the order in excess of 10^7 cd/m^2 while a starlit night in the range of 10^{-1} cd/m^2. This corresponds to more than 8 orders of magnitude. With regards to human eyes, their dynamic range depends on the time allowed to adjust or adapt to the given luminance levels. Due to the presence of rods and cones, human eyes have a remarkable ability to adjust to varying luminance levels, both dynamically (i.e. instantaneous) and over a period of time (i.e. adaptation time). Given sufficient adaptation time, the dynamic range of human eyes is about 13 orders of magnitude. However, without adaptation, the instantaneous human vision range is smaller and they are capable of dynamically adjusting so that a person can see about 5 orders of magnitude throughout the

entire range. Since the typical frequency in video signals does not allow sufficient adaptation time, the dynamic vision range (5 orders of magnitude) is more relevant in the context of this paper as well as HDR video processing in general. However, typical digital imaging sensors (assuming the typical single exposure setting) and LDR displays are not capable of dealing with such large dynamic range present in the real world, and most of them (both capturing sensors and displays) can handle upto 3 orders of magnitude. Due to this limitation, the scenes captured and viewed via LDR technologies will have lower contrast (visual details are either saturated or noisy) and smaller color gamut than what the eyes can perceive. This in turn can decrease the immersive experience quotient of the end-user.

HDR imaging technologies therefore aim to overcome the inadequacies of the LDR capture and display technologies via better video signal capture, representation and display, so that the dynamic range of the video can better match the instantaneous range of the eye. In particular, the major distinguishing factor of HDR imaging (in comparison to the traditional LDR one) is its focus on capturing and displaying scenes as natively (i.e. how they appear in the real world) as possible by considering physical luminance of the scene in question. Two important points should, however, be mentioned at the very outset. First, it may be emphasized that in HDR imaging one usually deals with proportional (and not absolute) luminance values. More specifically, unless there is a prior and accurate camera calibration, luminance values in an HDR video file represent the real world luminance upto an unknown scale[1]. This, nonetheless, is sufficient for most purposes. Secondly, the HDR displays currently available cannot display luminance beyond the specified limit, given the hardware limitations. This necessitates a pre-processing step for both subjective and objective HDR video quality measurement, as elaborated further in the next section. Despite the two mentioned caveats, HDR imaging can improve the viewer experience significantly as compared to LDR[2] imaging and, thus an active research area.

[1]Even with calibration, the HDR values represent real physical luminance with certain error. This is because the camera spectral sensitivity functions which relate scene radiance with captured RGB triplets cannot match the luminous efficiency function of the human visual system.

[2]The terms LDR and HDR are also sometimes respectively referred to as lower or standard dynamic range (SDR) and Higher Dynamic Range (to explicitly indicate that the range captured is only relatively higher than LDR but not the entire dynamic range present in a real scene). We, however, do away with such precise distinctions and always assume that the terms HDR and LDR are used in a relative context throughout this paper.

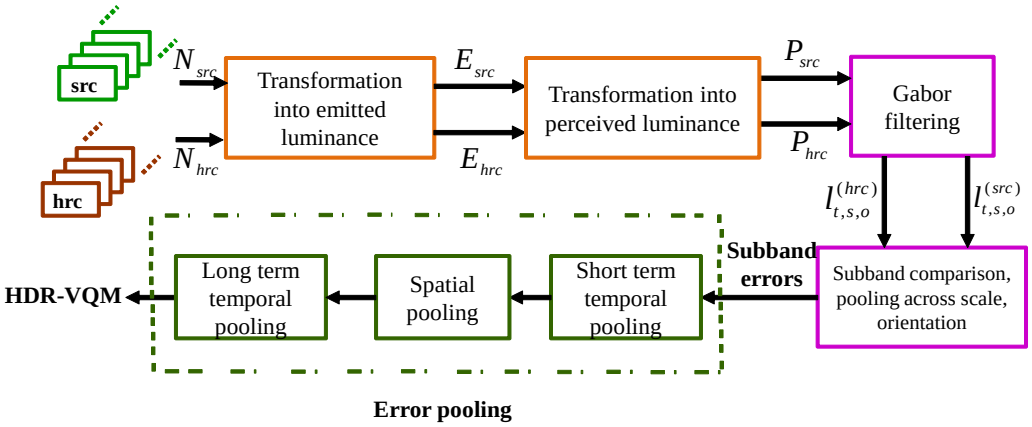


Fig. 1: Block diagram of the proposed HDR-VQM.

As already mentioned in the introduction, this paper seeks to address the issue of objective video quality measurement for HDR video. This is in light of the need to develop and validate such algorithms to objectively evaluate the perceptual impact of various HDR video processing tools on video quality. We describe the details of the method and verify its performance in the following sections.

III. The Proposed Objective HDR Video Quality Measure

A block diagram outlining the major steps in the proposed HDR-VQM is shown in Figure 1. It takes as input the source and the distorted HDR video sequences. Note that throughout the paper we use the notation src (source) and hrc (hypothetical reference circuit) to respectively denote reference and distorted video sequences. As shown in the figure, the first two steps are meant to convert the native input luminance to perceived luminance. These can therefore be seen as pre-processing steps. Next, the impact of distortions is analyzed by comparing the different frequency and orientation subbands in src and hrc. The last step is that of error pooling which is achieved via spatio-temporal processing of the subband errors. This comprises of short term temporal pooling, spatial pooling and finally, a long term pooling. A separate block diagram explaining the error pooling in HDR-VQM is shown in Figure 3. In the following sub-sections, we elaborate on the various steps in HDR-VQM.

A. From native HDR values to emitted luminance: modeling the display processing

We begin with two observations with regard to HDR video signal representation. First, as emphasized in the previous section, native HDR signal values are, in general, only proportional to the actual scene luminance and not equal to it. Therefore, the exact scene luminance at each pixel location will be, generally, unknown. Second, since the maximum luminance values of real-world scenes can be vastly different, the concept of a fixed maximum (or the white point) does not exist for HDR values. In view of these two observations, HDR video signals must be interpreted based on the display. Thus, their values should be re-calibrated according to the characteristics of the HDR display used to view them. This is unlike the case of LDR video where the format is more standardized (eg. for 8-bit representation, the maximum value will be 255 which would be mapped to the peak display luminance that does not typically exceed 500 cd/m^2). With regard to HDR displays, the inherent hardware limitations will impose a limit on the maximum luminance that can be displayed. Thus, a pre-processing of the HDR video signal is always required in order that the pre-defined maximum luminance point is not exceeded. Since this can affect distortion visibility, the said pre-processing for the transition from native HDR values (denoted as N_{src}, N_{hrc} in Figure 1), to the emitted luminance (denoted as E_{src}, E_{hrc} in Figure 1) is therefore an important step in the objective method design.

While one can adopt different strategies (from simple ones like linear scaling to more sophisticated ones) for the said display based pre-processing, this is not the main focus of the work. However, for the purpose of the method described in this paper, it is sufficient to highlight that in the general case, it is important that the characteristics of the HDR display are taken into account and the HDR video transformed (pre-processed) accordingly i.e. $N_{src} \rightarrow E_{src}$ and $N_{hrc} \rightarrow E_{hrc}$, for objective HDR video quality estimation. The specific pre-processing that we employed in our experiments in discussed in Section IV C.

B. Transformation from emitted to perceived luminance

The second step in the design of HDR-VQM concerns the transformation of the emitted luminance to perceived luminance i.e. $E_{src} \rightarrow P_{src}$ and $E_{hrc} \rightarrow P_{hrc}$ as indicated in Figure 1. This is required since there exists a non-linear relationship between the perceived and emitted luminance values given the response of the human visual system (HVS) to different luminance levels. An implication of such non-linearity is that the changes introduced by an HDR video

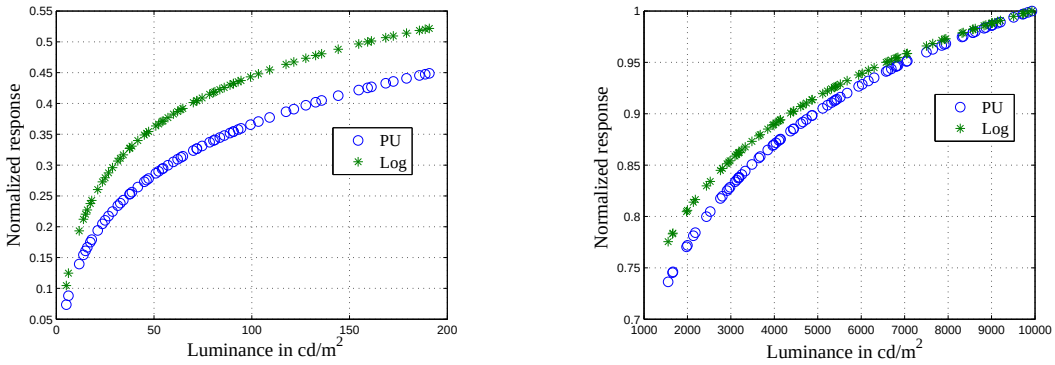


Fig. 2: Comparison of responses for two transformation functions in two different ranges of luminance, (a) 1 - 200 cd/m^2, (b) 1000 - 10000 cd/m^2. Figure best viewed in color.

processing algorithm in the emitted luminance may not have a direct correspondence to the actual modification of visual quality. This is different from the case of LDR representation in which the pixel values are typically gamma encoded. Thus, LDR video encodes information that is non-linearly (the non-linearity arising due to the gamma curve) related to the scene luminance. As a result of such non-linear representation, the changes in LDR pixel values can be approximately linearly related to the actual change perceived by the HVS. Due to this, many LDR image/video quality measurement methods directly employ the said gamma encoded pixel values as input and assume that changes in LDR pixels (or changes in features extracted from those pixels) due to distortion can quantify quality degradation (the reference video is always assumed to be of perfect quality). Therefore, to achieve a similar functionality as the LDR domain, the said non-linearity of the HVS to the emitted luminance should be taken into account for objective HDR video quality evaluation. In this way, the input values to the objective HDR video quality estimator would be expected to be approximately linearly related to the changes induced due to distortions.

To that end, a simple solution is to employ the logarithmic transformation which can approximate the saturation effect. However, the HVS is known to exhibit an approximate logarithmic response only for higher luminance (values in typically in the range 1000 - 10000 cd/m^2), and is more sensitive at lower luminance values [7]. We therefore explored the perceptually uniform

(PU) encoding proposed in [8] which can transform luminance values in the range from 10^{-5} to 10^{8} cd/m^2 into approximately perceptually uniform code values. In order to compare the two mentioned transformations, we plot in Figure 2, the respective responses to an input which is in the range from 1 - 10000 cd/m^2. For a closer look, we have plotted the responses for luminance in the range 1 - 200 cd/m^2 and above 1000 cd/m^2. We notice from Figure 2(a) that the response of PU encoding is relatively more linear at lower luminance as compared to the logarithmic one. To further quantify this, it was found that the linear correlation between the original and transformed signals was 0.9334 for PU encoding and 0.9071 for logarithmic, for the range between 1 - 200 cd/m^2. On the other hand, both PU and logarithmic curves have a similar response for higher luminance values (above 1000 cd/m^2) as indicated in Figure 2(b). In this case, the linear correlations were 0.8703 and 0.8763 respectively for PU and logarithmic transformation. Thus, PU encoding can better approximate the response of HVS which is approximately linear at lower luminance and increasingly logarithmic for higher luminance values. Due to this, PU encoding is expected to better model the underlying non-linear relationship between HVS's response and emitted luminance. The study on HDR video compression in [9] provides further evidence of this where it was demonstrated that PU encoding resulted in better coding efficiency as compared to the logarithmic transformation. From objective image quality assessment view point also, PU encoding has been shown [4] to be beneficial. Lastly, as described in the original paper [8], PU encoding has been implemented as a look-up table operation, and therefore does not substantially increase the computational overhead. All these features render PU encoding a reasonably accurate and efficient alternative for transforming the emitted luminance values to the perceived ones.

C. Computation of subband error signal

The proposed HDR-VQM is based on spatio-temporal analysis of an error video whose frames denote the localized perceptual error between a source and distorted video. We first describe the steps to obtain the subband error signal and then present the details of the spatio-temporal processing.

We employed log-Gabor filters, introduced in [10], to calculate the perceptual error at different scales and orientations. We note that log-Gabor filters have been widely used in image analysis and are a reasonable choice of features to compare intrinsic characteristics of natural scenes. In our approach, the log-Gabor filters were used in the frequency domain and can be defined

in polar coordinates by $H(f,\theta) = H_f \times H_\theta$ with H_f and H_θ being the radial and angular components, respectively:

$$H_{s,o}(f,\theta) = \exp(-\frac{\log(f/f_s)^2}{2(\log(\sigma_s/f_s)^2}) \times \exp(-\frac{(\theta-\theta_o)^2}{2\sigma_o^2}) \quad (1)$$

where $H_{s,o}$ is the filter denoted by spatial scale index s and orientation index o, f_s denotes the normalized center frequency of the scale, θ is the orientation, σ_s defines the radial bandwidth B in octaves with $B = 2\sqrt{2/log(2)} * |log(\sigma_s/f_s)|$, θ_o represents the center orientation of the filter, and σ_o defines the angular bandwidth $\Delta\Omega = 2\sigma_o\sqrt{2log(2)}$.

Video frames in the perceived luminance domain (i.e P_{src} and P_{hrc}) were decomposed into a set of subbands by computing the inverse DFT of the product of the frames's DFT with frequency domain filter defined in (1). We denote the resulting subband values as $\{l_{t,s,o}^{(src)}\}$ and $\{l_{t,s,o}^{(hrc)}\}$ respectively. Here, $s = 1, 2, ..., N_{scale}$ (the total number of scales), $o = 1, 2, ..., N_{orient}$ (the total number of orientations) and $t = 1, 2...F$ (the total number of frames in the sequence). The next step is to compute the error in the subband values for each frame at different scale and orientation levels. To that end, we chose the following simple and bounded measure (similar to the one used in [11]):

$$Error_{t,s,o} = \frac{2\ .\ l_{t,s,o}^{(src)}\ .\ l_{t,s,o}^{(hrc)} + k}{\{l_{t,s,o}^{(src)}\}^2 + \{l_{t,s,o}^{(hrc)}\}^2 + k} \quad (2)$$

Here k (a small constant) is added to avoid division by zero. We can then obtain the total error at each pixel in each video frame by pooling across scales and orientations. While more sophisticated methods such as those based on contrast sensitivity function (CSF) can be adopted for this purpose, a possible bottleneck is that of computing the desired CSF accurately (especially the one which may be applicable for both near-threshold and supra-threshold distortions). Therefore, in the current formulation, we limited the scope of application of the proposed HDR-VQM to supra-threshold distortions and chose to assign equal weighting as follows:

$$Error_t = \frac{1}{N_{scale} \times N_{orient}} \sum_{s=1}^{N_{scale}} \sum_{o=1}^{N_{orient}} Error_{t,s,o} \quad (3)$$

$Error_t$, which is a 2D distortion map, represents the error in each frame t due to processing of a source sequence by an HRC and the thus, the error video can be represented as $Error_{video} = \{Error_t\}_{t=1}^{F}$. It can be noted that the $Error_{video}$ helps to quantify the effect of local distortions by

assessing their impact across frequency and orientation. We can then exploit it further via spatio-temporal analysis in order to calculate short term quality in a spatially and temporally localized neighborhood, and subsequently obtain the overall HDR video quality score. We elaborate on this in the next sub-section.

D. From spatio-temporal subband errors to overall video quality: The Pooling step

Video signals propagate information along both spatial and temporal dimensions. However, due to visual acuity limitations of the eye, humans fixate their attention to local regions when viewing a video because only a small area of the eye retina, generally referred to as fovea, has a high visual acuity. As mentioned in Section II, this is due to higher density of photo receptor cells cones present in the fovea. Consequently, human eyes have to rapidly shift their gaze (the time between such movements is the fixation duration) to bring localized regions of the visual signal into the fovea field. Thus, humans tend to judge video quality in local context both spatially and temporally, and determine the overall video quality based on those assessments. Stated differently, the impact of distortions introduced in video frames will not be limited just to the spatial dimension but will rather manifest spatio-temporally. In the light of this, a reasonable strategy for objective video quality measurement is by analyzing the video in a spatio-temporal (ST) dimension [12], [13], [14], so that the impact of distortions can be localized along both spatial and temporal axes. To incorporate this in HDR-VQM, we first obtained the error video $Error_{video} = \{Error_t\}_{t=1}^{F}$ for the given pair of source and distorted HDR video. Then, the error video was divided into non-overlapping short-term ST tubes defined by a 3-dimensional region with x horizontal, y vertical and the z temporal data points i.e. a cuboid with dimensions $x \times y \times z$, as illustrated in Figure 3. Note that the former two define the spatial axes while the latter determines the temporal resolution to be considered when evaluating quality at the level of ST tubes. The values of x and y together define the area of the fixated region. Therefore, these can be computed by taking into account the viewing distance, the central angle of the visual field in the fovea and the display resolution. On the other hand, a good range of z can be determined by considering the average fixation duration when viewing a video. While this can vary due to content and/or distortions, studies (for eg. [15]) related to the analysis of eye-movement during video viewing indicate that values in the range of 300 - 500 ms can be a reasonable choice.

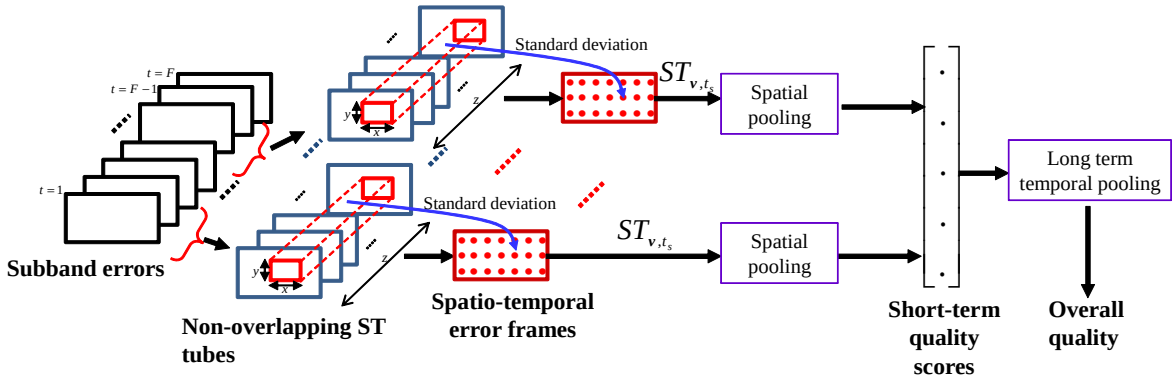


Fig. 3: Error pooling in HDR-VQM.

1) Short term temporal pooling: The aim of short term temporal pooling is to pool or fuse the data in local spatio-temporal neighborhoods (defined by ST regions). Keeping in mind that the goal is to characterize the effect of spatial distortions over short term duration (equal to fixation time), we computed the standard deviation of each ST tube in order to obtain a spatio-temporal subband error frame $ST_{\boldsymbol{v},t_s}$ (as visually exemplified in Figure 3) where $\boldsymbol{v}$ represents the spatial co-ordinates and t_s being the index of resulting spatio-temporal frames. By this definition, a video sequence with lower visual quality will have higher localized standard deviations while these will decrease as the signal quality improves. Thus, $ST_{\boldsymbol{v},t_s}$ help to quantify signal coherence level in local neighborhoods. Also note that in our case, the error video was analyzed based on non-overlapping ST tubes. Thus we will have $t_s = 1, 2, ..., F/z$. The final step is to perform spatial and long term temporal pooling to obtain the global video quality score.

2) Spatial and Long term temporal pooling: To obtain an overall video quality score that can quantify the level of annoyance in the video sequence, the local errors represented by spatio-temporal subband error frame $ST_{\boldsymbol{v},t_s}$ are pooled further in two stages: (a) spatial pooling to generate a time series of short term quality scores, (b) long term temporal pooling to fuse short term quality scores in to a single number denoting the overall annoyance level. These stages are based on the premise that humans evaluate the overall video quality based on continuous assessments of the impact of short term errors or annoyance they came across while viewing the video. Therefore, we first perform spatial pooling on spatio-temporal error frames $ST_{\boldsymbol{v},t_s}$ in order to obtain the short-term quality scores, as illustrated in Figure 3. We note that such short term

quality scores can be useful in providing information about the distortions temporally. Finally, a long term pooling is applied to compute the overall video quality score. We therefore use the following expression for HDR-VQM:

$$\text{HDR-VQM} = \frac{1}{|t_s \epsilon L_p| \times |\boldsymbol{v} \epsilon L_p|} \sum_{t_s \epsilon L_p} \sum_{\boldsymbol{v} \epsilon L_p} ST_{\boldsymbol{v},t_s} \tag{4}$$

where L_p denotes the set with lowest $p\%$ values, $|.|$ stands for cardinality of the set. The reader will notice that both short term spatial and long term temporal pooling have been performed over the lowest $p\%$ values. This is because the HVS does not process necessarily visual data in its entirety and makes certain choices to minimize the amount of data analyzed. It is, of course non-trivial to realize and integrate such exact HVS mechanisms into an objective method. In our case, we therefore employ the simple approach of percentile pooling keeping in mind that we are dealing with supra threshold distortions, and quality judgment would be based on how much the signal has been distorted (i.e. for a perfect quality video HDR-VQM will be zero). Finally, it is straightforward to use HDR-VQM to compute objective HDR image quality (obviously there will be no temporal index in this case).

IV. HDR VIDEO DATASET

To the best of our knowledge there are currently no publicly available subjectively annotated HDR video datasets dealing with the issue of visual quality. Therefore, for verifying the prediction performance of HDR-VQM and other objective methods, an in-house and comprehensive HDR video dataset was used. This section provides a brief description of the dataset.

A. Test material preparation

The dataset used 10 source HDR video sequences[3]. The frame resolution was full HD (1920 by 1080 pixels) and the frame rate was 25 frames per second (fps). The spatial versus temporal information measures (computed on tone mapped version of video frames) for each source sequence is shown in Figure 4. The source sequences were compressed at different bit rates to obtain the test stimuli. To that end, we employed a backward-compatible (in our context

[3]These HDR video sequences were shot and made available as part of the NEVEx project FUI11, related to HDR video chain study.

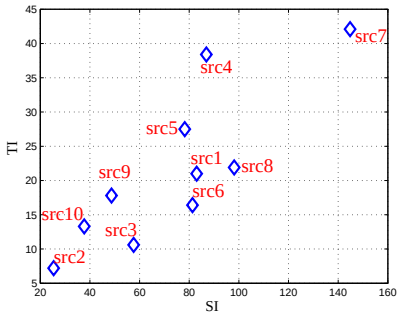


Fig. 4: Spatial and temporal information plot.

backward compatibility with existing standard 8-bit displays) HDR video compression method. In general, any backward-compatible HDR compression scheme comprises [16] of 3 steps: (a) forward tone mapping in order to convert HDR video to LDR (8-bit precision), (b) compression and decompression of the LDR video by a standard LDR video compression method, (c) inverse tone mapping of the decoded (decompressbed) LDR bit stream to reconstruct HDR video.The specific method that we used is described in [9]. It is not discussed here but the reader is encouraged to refer to [9] for details. The LDR video was encoded and decoded using H.264/AVC at different bit rates. For the subjective viewing tests, we selected 8 bit rates (by varying the quantization parameter QP) such that the resultant HDR video quality covered the entire rating scale, i.e., from excellent (rating 5) to bad (rating 1). With the inclusion of the source sequences, we obtained a total of 90 HDR video sequences (10×8 bit rates $+$ 10) to be rated by the subjects.

B. Rating methodology

Our study involved 25 paid observers who were not expert in image or video processing. They were seated in a standardized room conforming to the International Telecommunication Union Recommendation (ITU-R) BT500-11 recommendations [17]. Prior to the test, observers were screened for visual acuity by using a Monoyer optometric table and for normal color vision by using Ishiharas tables. All of them had normal or corrected to normal visual acuity and normal color perception. For rating the test stimuli, we adopted the absolute category rating with hidden reference (ACR-HR), which is one of the rating methods recommended by the ITU

in Rec. ITU-T P.910 [18]. The ACR-HR is a category judgment method where the test stimuli are presented one at a time and rated independently on a category scale. The rating method also includes the source sequences (i.e. undistorted) to be shown as any other test stimulus without informing the observers. This is therefore termed a hidden reference condition, and the advantage is that it implicitly encourages the observers to rate video quality and not fidelity since there is no reference to compare with. To quantify the video quality, a five-level scale is used: 5 (Excellent), 4 (Good), 3 (Fair), 2 (Poor) and 1 (Bad). We chose a discrete five-level scale because it is more suitable for naive (non-experts in image processing) observers, and it is easier for them to quantify the quality based on an adjective ('Excellent', 'Good', 'Fair', 'Poor' and 'Bad'). We also employed post-experiment screening of the subjects in order to reject any outliers in accordance with the Video Quality Experts Group (VQEG) multimedia test plan [19], and in our case no observer was rejected.

C. Display

For displaying the HDR video sequences, SIM2 Solar47 HDR display was used which has a maximum displayable luminance of 4000 cd/m^2. Since the HDR display luminance is relatively much higher as compared to the conventional display devices, we need much higher room illumination in order to reduce the visual discomfort of the subjects. In our study this was set to 200 cd/m^2 as it provided comfortable viewing conditions for the observers [20]. The viewing distance was set to three times the height of the screen (active part), i.e., approximately 178 cm.

It may be recalled from Section III A that we need to pre-process the HDR video signal in order that its luminance range fits that of the display i.e. $N_{src} \rightarrow E_{src}$ and $N_{hrc} \rightarrow E_{hrc}$. With regard to the said pre-processing step, the most straightforward way is the use of maximum normalization i.e. scaling the values in each video frame independently by the maximum luminance value in that frame. We however observed that this approach suffers from at least two drawbacks. First, because only a single luminance value (maximum) is used, it is susceptible to outliers. Second, we also observed that it tends to reduce the overall luminance coherence of the HDR video due to the local (i.e at frame level) normalization. To ameliorate these two issues, we opted for a temporally more coherent strategy and the normalization factor was determined as the maximum of the mean of top 5% luminance values of all the frames in an HDR video sequence. Specifically, we first computed a vector $\boldsymbol{MT_5}$ whose elements were the mean of top

5% luminance values in each HDR video frame i.e.

$$\boldsymbol{MT_5} = \left\{ \frac{1}{|\boldsymbol{v}\epsilon T_5|} \sum_{\boldsymbol{v}\epsilon T_5} N_{\boldsymbol{v},t} \right\}_{t=1,2...F} \tag{5}$$

where N denotes the native HDR values at spatial location $\boldsymbol{v}$ for frame t (F is the total number of frames), T_5 denotes the set with highest 5% luminance values in the frame. Then, the native HDR values N were converted to emitted luminance values E as

$$E = \frac{N \times 179}{max(\boldsymbol{MT_5})} \tag{6}$$

where the multiplication factor of 179 is the luminous efficacy of equal energy white light that is defined and used by the Radiance file format (RGBE) for the conversion to actual luminance value. Finally, a clipping function was applied to limit the E values in the range defined by the black point (lowest displayable luminance) and the maximum displayable luminance (obviously both will depend on the display characteristics).

V. Experimental Results

Before we compute the experimental results, it is necessary to determine the parameter values in HDR-VQM. First, to compute the values of x and y, we assume the central angle of the visual field in the fovea to be 2° [21]. Then, we can define a quantity W which denotes the length of the fixated window in terms of number of pixels and can be computed as

$$W = \tan 2^\circ \times V \times \sqrt{R/D_A} \tag{7}$$

where V is the viewing distance in cm., R is the display resolution and D_A is the display area. In our case, $V = 178$ cm, $R = 1080 \times 1920$ pixels and $D_A \approx 6100$ cm^2. Plugging these values in to (6) gives $W \approx 115$. To reduce the computational effort, HDR-VQM was run on down sampled (by a factor of 2) video frames, and hence the approximate length of the fixated window will be $W/2 \approx 58$. Thus, we set $x = y = 64$ pixels in order to be nearest to a more standard block size. To determine z, we assumed a fixation duration of 400 ms and with a frame rate of 25 fps, we will have $z = 10$ frames. The number of scales and orientations used were 5 and 4, respectively i.e. $N_{scale} = 5$ and $N_{orient} = 4$.

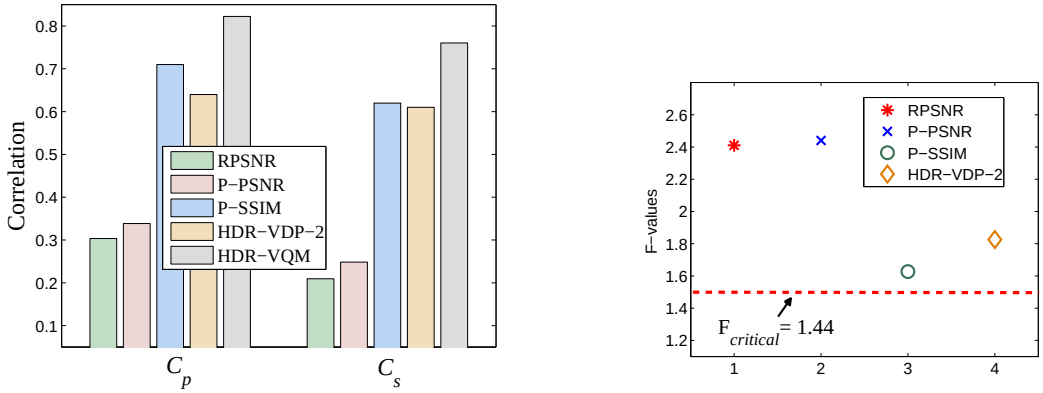
A. Correlation based comparisons

The first set of experimental results are reported in terms of two criteria: Pearson linear correlation coefficient C_p (for prediction accuracy) and Spearman rank order correlation coefficient C_s (for monotonicity), between the subjective score and the objective prediction. The former depends on the non-linear mapping between objective and subjective scores (we therefore employed a four-parameter monotonic logistic mapping between the objective outputs and the subjective quality ratings [19]) while the latter is independent of any such non-linearity. We compared the performance of HDR-VQM with a few popular LDR methods including PSNR and multi-scale SSIM [11]. The input to all these methods were the perceived luminance values P_{src} and P_{hrc} and hence we refer to them as P-PSNR and P-SSIM. In addition, we considered two other methods which take as input the emitted luminance values E_{src} and E_{hrc}. The first one is the HDR-VDP-2 originally proposed in [22], and we employed its re-calibrated version reported in [23]. The second method is the relative PSNR (RPSNR) which employs a variant of the traditional mean squared error, MSE, in that it normalizes the error by the magnitude of luminance at each point in order to account for reduced sensitivity at high luminance. It can be defined as

$$RPSNR = -10\ log_{10}\Big(\ \frac{1}{N_{pixels}} \sum_{i,j} \frac{(E_{src_{i,j}} - E_{hrc_{i,j}})^2}{E^2_{src_{i,j}} + E^2_{hrc_{i,j}}}\Big) \tag{8}$$

where i, j and N_{pixels} respectively denote the pixel index and the total number of pixels in the source and distorted frames.

Since we are dealing with full-reference video quality estimation, we computed the correlations using only the 80 distorted sequences (i.e. 80 test conditions), and these are shown in Figure 5. The results for across sequences (i.e. with all the 80 sequences) are shown in Figure 5(a). One can observe the HDR-VQM achieves relatively higher prediction accuracy in comparison to the other methods. It is also interesting to point out the proposed HDR-VQM, P-PSNR, and P-SSIM compute quality based on perceived luminance. The better performance of HDR-VQM relative to these methods therefore indicates the added value of taking into account frequency and orientation information. We also carried out analysis based on an F-test to investigate the significance of the obtained results from a statistical view point, and those results are represented graphically in Figure 5(b). In this, the F values were computed for the different methods with respect to HDR-VQM (the residuals were verified to be approximately Gaussian which is an



(a) (b)

Fig. 5: Numerical results, (a) C_p and C_s for all 80 distorted sequences, (b) F-test results. Figure best viewed in color.

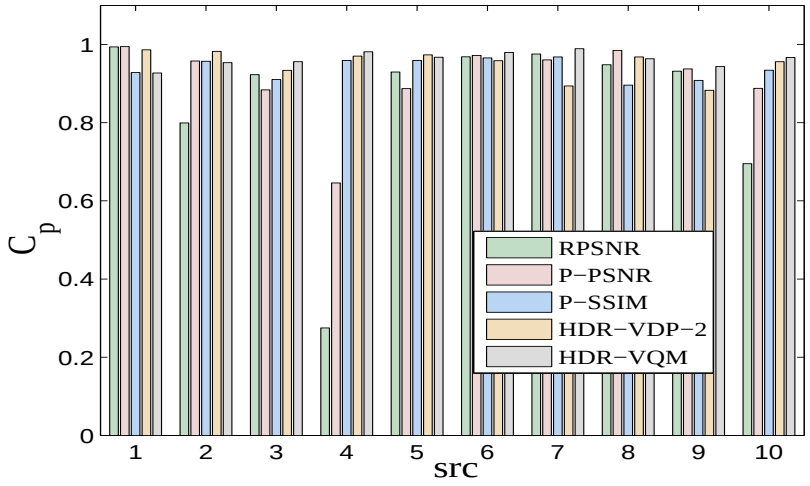


Fig. 6: C_p values for per source (src) sequence (these values are computed without the logistic fitting function). Figure best viewed in color.

assumption for F-test). The cases for which $F > F_{critical}$ (i.e. points above the $F_{critical}$ line) indicate that HDR-VQM is significantly better than the corresponding method. Of course, a larger test bed could help to further bring out such statistical differences between objective methods.

Finally, the results, i.e. C_p values (without any non-linear fitting since the number of data

points to be fitted is small), for per source sequence are indicated in Figure 6 from which we note that simpler methods like P-PSNR and MRSE tend to perform better (although the performance is still not consistent for each source) as compared to across sequence. But, in general, HDR-VQM is competitive or better in majority of cases.

To further explain why simpler methods like RPSNR and P-PSNR achieve relatively lower correlations in case of across sequence in comparison to the case of per sequence, the scatter plots for different methods are shown in Figure 7. In this, the x-axis denotes the objective scores while the y-axis indicates the MOS. We can see for instance in Figure 7(a) that while the RPSNR values per source are more linearly related to the MOS, they exhibit a large scatter due to the curves being shifted for each source. As a result, for similar MOS values across sequences, the corresponding RPSNR values can be quite different. This decreases the overall correlations. On the other hand, the HDR-VQM, while not perfect, exhibits less scatter across sequences indicating its ability to better distinguish between quality levels of video sequences from across different sources. To illustrate this further, we take two specific conditions for src1 and src6. One can see that for RPSNR achieves quite high correlations (in both these cases it is higher than 0.96) for both these source sequences. Consider the conditions: (a) src1 encoded at the lowest bit-rate (highest QP), (b) src6 encoded at the highest bit-rate (lowest QP). The RPSNR value for the first condition was 32.70 dB while the corresponding subjective score was 1.04. On the other hand, for the second condition, the RPSNR was 28.54 dB with the subjective score for this condition being 3.92. Thus, just looking at RPSNR values can lead to the conclusion that the HDR video in first condition has higher quality than the one in the second condition, which is not the case according to subjective opinion. On the other hand, the HDR-VQM values for the two conditions (respectively 0.1795 and 0.01, recall smaller HDR-VQM implies higher quality) are more in line with subjective opinion. We also wish to stress that this example is only meant to highlight specific potential problems with simpler methods like RPSNR. Of course, one should rely on correlation based comparisons and outlier ratio analysis (presented in the next subsection) to draw more general conclusions about the performance of different objective methods.

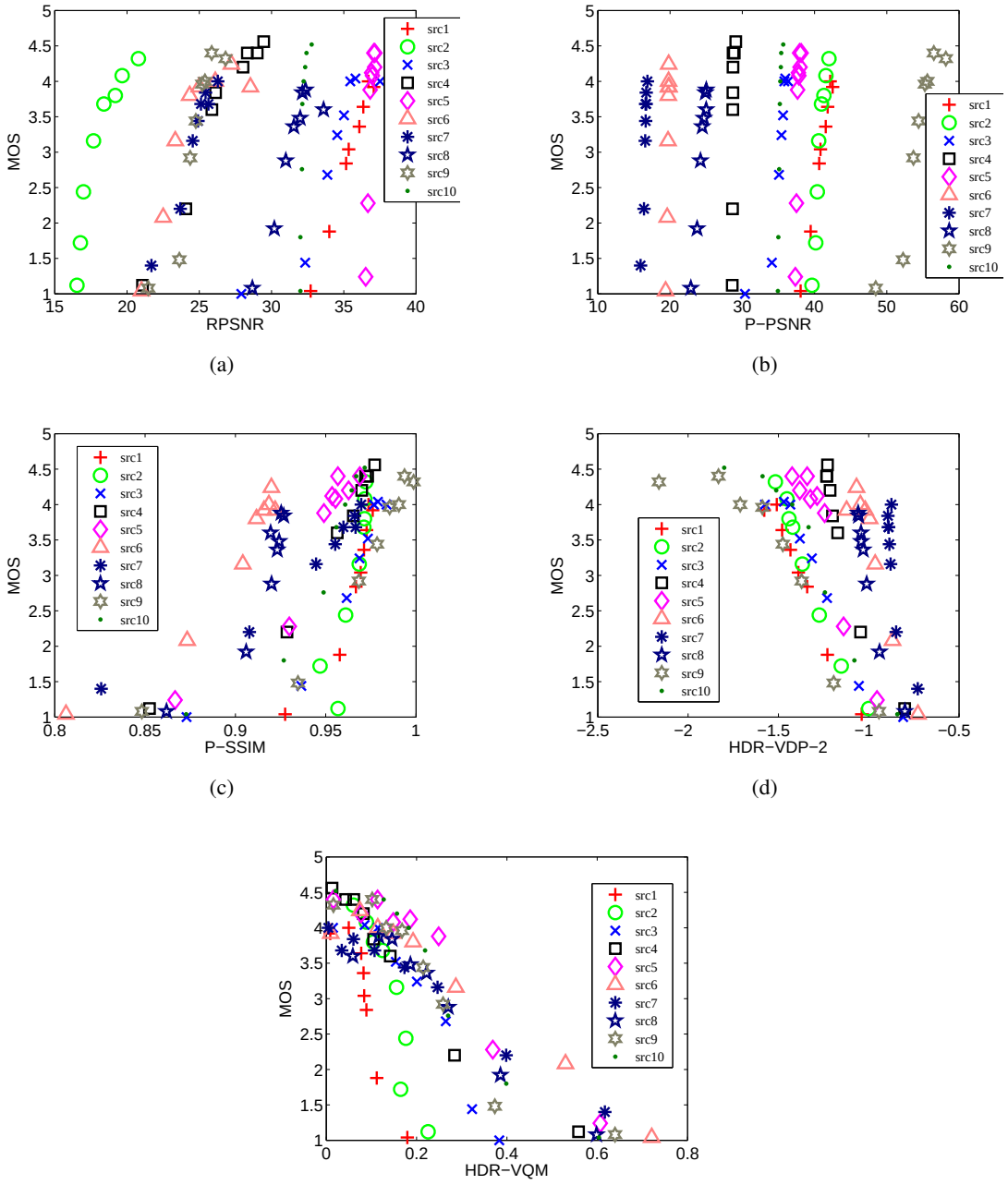


Fig. 7: Scatter plots for different objective methods, (a) RPSNR, (b) P-PSNR, (c) P-SSIM, (d) HDR-VDP-2 and (e) HDR-VQM. Figure best viewed in color.

TABLE I: Outlier ratio analysis for the 80 test conditions in the HDR video dataset.

	R-PSNR	P-PSNR	P-SSIM	HDR-VDP-2	HDR-VQM
Number of Outliers	42	46	35	38	18
% of Outliers	52%	57%	43%	47%	22%

B. Outlier ratio analysis

Outlier ratio analysis is another approach to evaluate objective methods for their prediction accuracy. It is different from the usual correlation based comparisons in that it does not penalize if the prediction error is within the limit of uncertainty in the subjective rating process. Particularly, it can be very useful in applications such as video compression where one is generally interested in the rate distortion (RD) behavior of objective methods i.e. how the objective visual quality varies with bit rates for different source sequences and to what extent that compares with the subjective video quality. To carry out the outlier ratio analysis, we computed the number of outliers from different objective methods for the 80 conditions used in our HDR video dataset. To define an outlier, we use the fact that the subjective rating process is always associated with uncertainty due to inter-observer differences between the perceived quality of the same stimuli. Consequently, the actual subjective video quality can vary between two limits which can be, for instance, determined based on confidence intervals. Therefore, we first computed the absolute prediction error between the subjective MOS and logistically transformed objective scores for each of the 80 test conditions. For each objective method, if the said error was greater than twice the 95% confidence interval associated with the MOS, then that objective method was deemed as an outlier for the given condition. Formally, if $|MOS_c - OM_c^{(logistic)}| > 2CI_c$ then the objective prediction from the given objective method OM will be deemed as an outlier for condition c ($c = 1, 2, ...80$). Here MOS_c, $OM_c^{(logistic)}$ and CI_c respectively denote the subjective score, the logistically transformed value from the method OM and the associated 95% confidence interval, for condition c. The resulting number of outliers for each objective method is shown in Table I along the percentage of outliers for the 80 test conditions.

We find that HDR-VQM has the least number of outliers (22%). Also notice that P-SSIM, which was second in terms of correlation results in Figure 5(a), has about 43% outliers which is almost double that of HDR-VQM. Thus, such outlier analysis provides additional insights

into the behavior of different objective methods which is probably less apparent looking at just the correlation values. The main advantage of outlier analysis is that it helps to evaluate metric accuracy by taking into account the variability or uncertainty (expressed via 95% confidence intervals in our dataset) in subjective opinions, which are ignored in correlation based comparisons. Overall we find that HDR-VQM is reasonably accurate and more consistent for both across and per sequence cases as indicated by different tests (including correlation coefficients, statistical test and outlier analysis).

VI. Discussion

The previous sections proposed and verified the performance of an objective HDR video quality estimator HDR-VQM. An important point worth re-iterating is that in light of limitations imposed in HDR display technologies, the HDR video signal must be pre-processed. This, in turn implies that HDR video quality can be affected by the type of pre-processing used as well as the maximum displayable luminance since these can affect distortion visibility. Thus, the role of a display model is more prominent in HDR, and objective quality assessment should consider this aspect. As opposed to this, quality measurement for the traditional LDR video signals does not usually require any specific display based pre-processing. By this, we do not imply that the display does not affect quality judgment in LDR domain, merely its effect is perhaps less prominent.

While the proposed HDR-VQM is reasonably accurate, it is certainly not without its limitations as is the case with any objective method. The reader will notice that HDR-VQM does not explicitly account for spatio-temporal masking that could play a role in quality judgment. Therefore, in the current form, it is assumed that the employed measure of similarity is directly related to the change in video quality. We note that such an assumption is more reasonable for supra-threshold distortions (i.e. distortions clearly visible to the human eye) but may not always hold in case of near-threshold distortion levels where the impact of spatio-temporal masking may be more prominent. Also recall that in (3) we did not employ a more sophisticated weighting such as one based on CSF. Thus, in the current formulation, HDR-VQM may be more effective in case of supra-threshold distortions and we believe that it is important to make this distinction in order to define the scope of application.

With regard to the strengths of the proposed method, one of the most prominent is its relatively

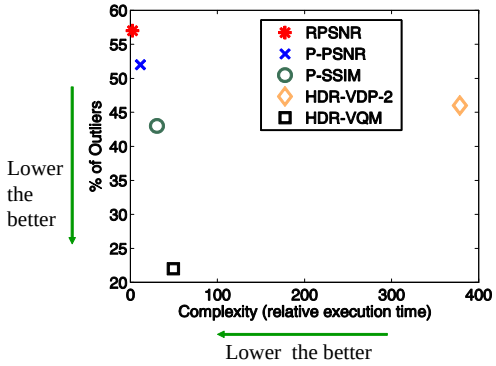


Fig. 8: % of outliers and the relative computational complexity (expressed as relative execution time with respect to RPSNR) for different methods. Figure best viewed in color.

low computational complexity due to the fact that it does not explicitly extract motion information from the video. Rather, the focus is on exploiting basic video signal characteristics by resorting to spatio-temporal analysis which, in turn, is based on the idea that humans tend to judge video quality in local context, both from spatial and temporal view points. As a consequence, HDR-VQM also provides certain local spatio-temporal quality information that can be for instance useful as feedback for video codec optimization. While a formal computational complexity analysis of HDR-VQM is not the focus of this work, recall that in HDR-VQM we performed the filtering in the frequency domain. This allows us to exploit the efficiency of the Fast Fourier Transform (FFT) and inverse FFT whose complexity is $O(M^2 log M)$ where M is the frame size. Further, Figure 8 provides the percentage of outliers and the relative computational complexity (expressed as the execution time realtive to RPSNR). Obviously, lower values for an objective method along both axes implies that it is better. We find that the relative execution time for HDR-VQM is reasonable considering the improvements (i.e. smallest % of outliers) in performance over other methods. The reader will also appreciate the fact that video quality judgment, in general, can depend on several extraneous factors (such as display type, viewing distance, ambient lighting conditions etc.) apart from the distortions themselves. In that regard, a notable feature of HDR-VQM is that it takes into account some of the mentioned factors such as the viewing distance, fixation duration and display resolution in computing the different parameters. This allows HDR-VQM to adapt to some of the physical factors that may affect subjective quality

judgment. Finally, HDR-VQM does not have any parameter that needs to be tuned/trained and hence does not require a priori information about the content or distortion.

VII. Concluding Thoughts

HDR imaging is increasingly becoming popular in the multimedia signal processing community primarily as a tool towards enhancing the immersive video experience of the user. However, there are very few works that address the issue of assessing the impact of HDR video processing algorithms on the perceptual video quality both from subjective and objective angles. The study in this paper seeks to outline the first steps towards the design and verification of an objective HDR video quality estimator HDR-VQM. While objective video quality measurements cannot always replace subjective opinion, they can still be useful in specific HDR video processing applications where subjective studies may be infeasible (eg. real-time video streaming, video encoding etc.). To that extent and within the scope of its application, HDR-VQM is a reasonable objective tool for HDR video quality measurement. To enable others to use the proposed method as well as validate it independently, a software implementation will soon be made available online for free download and use. The immediate future work will ensue further refinement of the presented method in view of some of the mentioned limitations as well as further validation with larger HDR video datasets.

Acknowledgment

The authors wish to thank Romuald Pepion for his help in generating the subjective test results used in this paper. This work has been supported by the NEVEx project FUI11 financed by the French government.

References

[1] P. L. Callet, S. Moller, and A. Perkis, "Qualinet white paper on definitions of quality of experience (2012)," White Paper, European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003), 2013.

[2] F. Banterle, A. Artusi, K. Debattista, and A. Chalmers, *Advanced High Dynamic Range Imaging: Theory and Practice*. Natrick, MA, USA: AK Peters (CRC Press), 2011.

[3] M. Narwaria, M. Silva, P. L. Callet, and R. Pepion, "Tone mapping-based high-dynamic-range image compression: study of optimization criterion and perceptual quality," *Optical Engineering*, vol. 52, no. 10, pp. 102 008–102 008, 2013.

[4] G. Valenzise, F. Simone, P. Lauga, and F. Dufaux, "Performance evaluation of objective quality metrics for hdr image compression," in *Proc. SPIE*, vol. 9217, 2014, pp. 92 170C–92 170C–10.

[5] M. Narwaria, M. Silva, P. L. Callet, and R. Pepion, "On improving the pooling in HDR-VDP-2 towards better hdr perceptual quality assessment," in *Proc. SPIE*, vol. 9014, 2014, pp. 90 140N–90 140N–9.

[6] G. Mather, *Foundations of Perception*. Psychology Press, Hove, East Sussex, 2006.

[7] R. Mantiuk, K. Myszkowski, and H. Seidel, "Lossy compression of high dynamic range images and video," in *Proc. SPIE*, vol. 6057, 2006, pp. 60 570V–60 570V–10.

[8] T. Aydin, R. Mantiuk, and H. Seidel, "Extending quality metrics to full luminance range images," in *Proc. SPIE*, vol. 6806, 2008, pp. 68 060B–68 060B–10.

[9] A. Koz and F. Dufaux, "Optimized tone mapping with perceptually uniform luminance values for backward-compatible high dynamic range video compression," in *Proc. IEEE Visual Communications and Image Processing (VCIP)*, Nov 2012, pp. 1–6.

[10] D. Field, "Relations between the statistics of natural images and the response properties of cortical cells," *J. Opt. Soc. Am. A*, vol. 4, no. 12, pp. 2379–2394, Dec 1987. [Online]. Available: http://josaa.osa.org/abstract.cfm?URI=josaa-4-12-2379

[11] Z. Wang, E. Simoncelli, and A. Bovik, "Multiscale structural similarity for image quality assessment," in *Proc. IEEE Conference Record of the Thirty-Seventh Asilomar Conference on Signals, Systems and Computers*, vol. 2, Nov 2003, pp. 1398–1402.

[12] M. Pinson and S. Wolf, "A new standardized method for objectively measuring video quality," *IEEE Trans. Broadcast.*, vol. 50, no. 3, pp. 312–322, Sept 2004.

[13] M. Narwaria, W. Lin, and A. Liu, "Low-complexity video quality assessment using temporal quality variations," *IEEE Trans. Multimedia*, vol. 14, no. 3, pp. 525–535, June 2012.

[14] A. Ninassi, O. Meur, P. L. Callet, and D. Barba, "Considering temporal variations of spatial visual distortions in video quality assessment," *IEEE J. Sel. Topics Signal Process.*, vol. 3, no. 2, pp. 253–265, April 2009.

[15] O. Meur, A. Ninassi, P. L. Callet, and D. Barba, "Overt visual attention for free-viewing and quality assessment tasks: Impact of the regions of interest on a video quality metric," *Signal Processing: Image Communication*, vol. 25, no. 7, pp. 547 – 558, 2010.

[16] M. Narwaria, M. Silva, and P. L. Callet, "High dynamic range visual quality of experience measurement: Challenges and perspectives," in *Visual Signal Quality Assessment*, C. Deng, L. Ma, W. Lin, and K. N. Ngan, Eds. Springer International Publishing, 2015, pp. 129–155. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-10368-6_5

[17] "Methodology for the subjective assessment of the quality of television pictures." Recommendation ITU-R BT.500-13, Jan 2012. [Online]. Available: http://www.itu.int/rec/R-REC-BT.500-13-201201-I

[18] "Subjective video quality assessment methods for multimedia applications." ITU-T Recommendation P.910, April 2008. [Online]. Available: http://www.itu.int/rec/T-REC-P.910-200804-I

[19] "Final report from the video quality experts group on the validation of objective quality metrics for video quality assessment." Video Quality Experts Group (VQEG), March 2003.

[20] "Tone mapping based HDR compression: Does it affect visual experience?" *Signal Processing: Image Communication*, vol. 29, no. 2, pp. 257 – 273, 2014.

[21] D. Green, "Regional variations in the visual acuity for interference fringes on the retina," *J Physiol.*, vol. 207, no. 2, pp. 351 – 356, Apr 1970.

[22] R. Mantiuk, K. Kim, A. Rempel, and W. Heidrich, "HDR-VDP-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions," *ACM Trans. Graph.*, vol. 30, no. 4, pp. 40:1–40:14, Jul. 2011.

[23] M. Narwaria, R. Mantiuk, M. Silva, and P. L. Callet, "Calibrated HDR-VDP-2 for objective quality assessment of high dynamic range and standard images," *Journal of Electronic Imaging*, Accepted (pending minor revisions).